

Soundex – Reference Guide

Version 1.0, December 1999

<u>Topic</u>	<u>Page</u>
Summary: Soundex	1
Definition	1
Data Storage and Field Values	1
Missing Values	2
Processing Overview	2
EDI Summary	2
Discussion	2
Implementation: Soundex	3
Data Collection: Hardcopy Report Form	3
Data Entry: Electronic Forms	4
Data Processing: Validations and Edit Checks	6
Data Processing: From Hardcopy to Storage	9
Data Transmission: Electronic Data Interchange	10

Summary: Soundex**Definition**

Soundex code for person's last name.

Data Storage and Field Values

There is 1 data element used to define the concept Soundex. The variable name noted below specifically applies to the soundex associated with the surname of the subject of the report. Variable names for other uses of soundex, such as soundex associated with the surname of a contact of the subject, are not discussed in this document but will be discussed in subsequent releases.

Soundex

Description:	Soundex code for person's last name
Variable Name:	SOUNDEX
Type:	character
Length:	4
Reported to CDC:	Yes
Field Values:	Character 1: A-Z Characters 2-4: 0-9

Missing Values

If the value of the Soundex data element is missing, or does not adhere to the CIPHER standard, the data element may be noted as blank to indicate a missing value. If the program requires the reason the value is missing, use a separate 1-character field to denote the rationale behind the missing data. The use of a Missing Value Reason data element must adhere to the CIPHER definition and rules associated with missing data as described in Appendix I - Missing Value Reason.

Processing Overview

Special requirements apply. Refer to the Implementation subsection on Data Processing: Validations and Edit Checks, below, for detailed information.

EDI Summary

EDI Sections are under construction.

Discussion

Soundex is a phonetic, alphanumeric code created by converting a surname into an index letter and a 3-digit code. The index letter is the first letter of the surname. The 3-digit code is calculated from the remaining letters of the surname, according to the set of rules described in the subsection on Data Processing: Validations and Edit Checks (below).

The advantage of Soundex is its ability to group names by sound rather than by exact spelling. For reasons of confidentiality, person name data are not transmitted to CDC. However, Soundex data are permitted for transfer to CDC because a person name cannot be inferred from a phonetic Soundex code. By using Soundex in conjunction with other demographic data such as date of birth and sex, CDC programs are able to identify duplicate name reports while adhering to patient confidentiality and privacy laws.

The Soundex code is typically calculated by automatic application of a soundex algorithm to the person's surname. The Soundex code can also be entered manually when obtained from or calculated by an external source. In either case, the Soundex code is stored in the 4-character SOUNDEX variable.

Implementation: Soundex

The implementation examples noted below specifically apply to the Soundex associated with the surname of the subject of the report. The implementation for other uses of Soundex, such as Soundex associated with the surname of a contact of the subject of the report, can be patterned after these implementation examples.

Data Collection: Hardcopy Report Form

The Soundex code is typically calculated by automatic application of a Soundex algorithm to the surname. However, programs may choose to manually collect/enter a Soundex code in situations in which the Soundex code is obtained or calculated from an external source. In these instances, a free-form entry field on the hardcopy report form is used for the collection of Soundex data. Refer to Figures 1 and 2 below.

Figure 1: **Blank Hardcopy Form section used to collect Soundex**

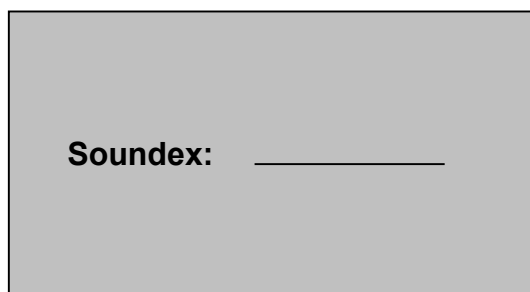
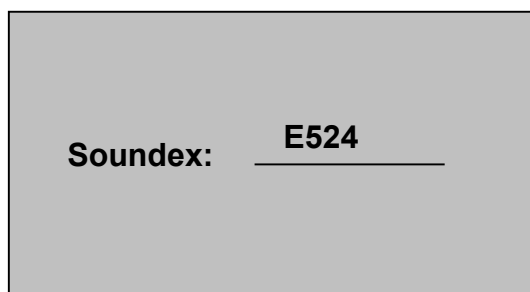
A rectangular box with a light gray background. Inside the box, the text "Soundex:" is followed by a horizontal line, indicating a space for manual entry.

Figure 2: **Completed Hardcopy Form section used to collect Soundex**

A rectangular box with a light gray background. Inside the box, the text "Soundex:" is followed by a horizontal line, and the code "E524" is entered on the line.

Missing Values – Hardcopy Form

Examples of hardcopy forms using the associated Missing Value Reason data element can be found in Appendix I – Missing Value Reason. The hardcopy form need only contain a missing value reason if the program requires the rationale for a missing value for Soundex.

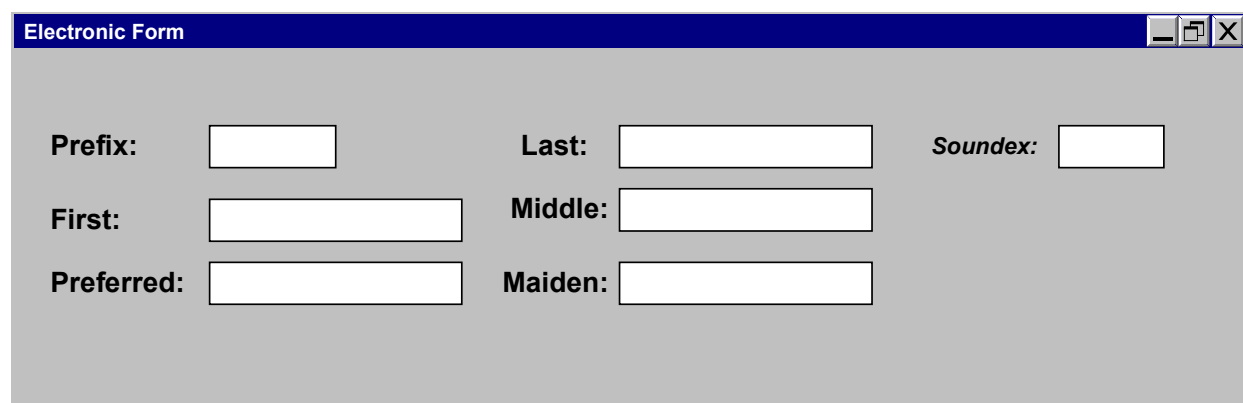
Data Entry: Electronic Forms

The Soundex entry field is typically populated (automatically filled) with a Soundex code following the entry of the associated surname. That is, the soundex algorithm is automatically applied to the surname, resulting in the generation of a Soundex code, which is then populated in the Soundex field. Refer to Figures 3 and 4 for an illustration.

In addition, the Soundex code can be entered manually when the entry of surname is bypassed. In these instances, the Soundex code is obtained or calculated from an external source. Refer to Figure 5 for an illustration.

Because the Soundex code is closely tied to the surname data, the following illustrations contain fields that define the Person Name data concept as well as the Soundex code field itself.

Figure 3: Blank Electronic Form used to collect Soundex



The image shows a screenshot of a software window titled "Electronic Form". The window has a blue title bar with standard minimize, maximize, and close buttons. The main area is light gray and contains several text input fields arranged in three rows. The first row has "Prefix:" followed by a small text box, "Last:" followed by a medium text box, and "Soundex:" followed by a small text box. The second row has "First:" followed by a medium text box and "Middle:" followed by a medium text box. The third row has "Preferred:" followed by a medium text box and "Maiden:" followed by a medium text box.

Prefix:	<input type="text"/>	Last:	<input type="text"/>	Soundex:	<input type="text"/>
First:	<input type="text"/>	Middle:	<input type="text"/>		
Preferred:	<input type="text"/>	Maiden:	<input type="text"/>		

Figure 4: Completed Electronic Form, with surname data entered, and Soundex automatically calculated following the entry of the surname data

The screenshot shows a window titled "Electronic Form" with a grey background. It contains several text input fields arranged in two columns. The first column has labels "Prefix:", "First:", and "Preferred:". The second column has labels "Last:", "Middle:", and "Maiden:". The "Soundex:" label is positioned to the right of the "Last:" field. The "Last:" field contains the text "ANDERSON", and the "Soundex:" field contains "A536". The other fields are filled with: "Prefix:" is "Mr.", "First:" is "JACKSON", "Middle:" is "ROBERT", "Preferred:" is "JACK", and "Maiden:" is empty. A callout box with a black border and white background is located to the right of the "Soundex:" field. It contains the text "Automatically calculated upon entry of the last name" with a black arrow pointing from the text to the "Soundex:" field.

Prefix:	<input type="text" value="Mr."/>	Last:	<input type="text" value="ANDERSON"/>	Soundex:	<input type="text" value="A536"/>
First:	<input type="text" value="JACKSON"/>	Middle:	<input type="text" value="ROBERT"/>		
Preferred:	<input type="text" value="JACK"/>	Maiden:	<input type="text"/>		

Automatically calculated upon entry of the last name

Figure 5: Completed Electronic Form, with surname data bypassed, and Soundex manually entered.

The screenshot shows a window titled "Electronic Form" with a grey background. It contains several text input fields arranged in two columns. The first column has labels "Prefix:", "First:", and "Preferred:". The second column has labels "Last:", "Middle:", and "Maiden:". The "Soundex:" label is positioned to the right of the "Last:" field. The "Last:" field is empty, and the "Soundex:" field contains "A536". The other fields are empty: "Prefix:", "First:", "Preferred:", "Middle:", and "Maiden:".

Prefix:	<input type="text"/>	Last:	<input type="text"/>	Soundex:	<input type="text" value="A536"/>
First:	<input type="text"/>	Middle:	<input type="text"/>		
Preferred:	<input type="text"/>	Maiden:	<input type="text"/>		

Missing Values – Electronic Form

Examples of electronic forms using the associated Missing Value Reason (MVR) data element can be found in Appendix I – Missing Value Reason. The electronic form needs to handle the Missing Value Reason only if the program requires the rationale for a missing value for Soundex. If the user selects a missing value reason code during data entry, the Soundex field will be blank and the screen will display the MVR information next to the blank field.

Data Processing: Validations and Edit Checks

Data elements entered in the electronic form will be edited as outlined below. If the program elects to use an associated Missing Value Reason data element for Soundex, it will be edited as outlined in Appendix I – Missing Value Reason.

Soundex is a phonetic, alphanumeric code calculated by converting a surname into a 4-digit format consisting of an index letter and a 3-digit code. The index letter is the first letter of the surname. As noted below, Character 1 contains the first letter of the surname, and Characters 2-4 contain a 3-digit code calculated from the remaining letters of the surname.

Character 1	Contains the index letter, which is the first letter of the surname. Valid values range from A-Z.
Characters 2 - 4	Contains a 3-digit code calculated from the remaining letters of the surname according to a set of rules. Valid values for each of the three digits range from 0-9.

The soundex code is calculated from the surname according to the following rules:

1. The first letter of the surname is used in the first position (Character 1).
2. The vowels A, E, I, O, U, and Y and the consonants H and W are never coded.
3. The remaining consonants of the surname are represented by key letters or their letter (phonetic) equivalents. Each is assigned a numeric code as shown below:

<u>Key</u>	<u>Letter Equivalents</u>	<u>Numeric code</u>
B	F, P, V	1
C	G, J, K, Q, S, X, Z	2
D	T	3
L		4
M	N	5
R		6

4. Letters are converted into numeric codes in the order in which they appear. Consider the following examples:

H	O	L	M	E	S	H452
		4	5		2	

G	W	I	L	F	O	Y	L	E	G414
			4	1				4	

5. Codes always contain 3 digits, and no more than 3 digits. Codes for names that do not contain 3 key letters or their equivalents are completed by adding zeros.

G R A H A M	G650
6 5	

B A I L E Y	B400
4	

S H A W	S000
---------	------

NOTE: Zeros follow, but never precede, the numeric codes 1-9.

6. Codes for names that contain more than 3 key letters or their equivalents use only the first 3 digits:

V O N D E R L E H R	V536
5 3 6 - -	

7. Two or more consecutive key letters or their equivalents are treated as one key letter and are coded as one digit:

B A L L O U	B400
4 -	

J A C K S O N	J250
2 - - 5	

8. Key letters or their equivalents that follow an initial letter from the same *phonetic* (letter equivalent) group are not coded:

S C A N L O N	S545
- 5 4 5	

S C K L A R	S460
- - 4 6	

9. Key letters or their equivalents separated by A, E, I, O, U, or Y are coded separately:

H A N N O N	H550
5 - 5	

S A L K I E W I C Z	S422
4 2 2 -	

10. Abbreviated prefixes, such as “Mc” and “St.” are coded as if spelled out:

McIlhaney:
M A C I L H A N E Y M245
2 4 5

St. John:
S A I N T J O H N S532
5 3 2 -

11. Apostrophes are disregarded:

O ' N E I L O540
5 4

Data Processing: From Hardcopy to Storage

The following example illustrates the flow of information from data collection on the hardcopy form, to data entry into the electronic form, to validations and storage in the database.

The process begins with the blank Hardcopy data collection form used to collect Soundex:



In some instances, the Soundex information is captured on the form, creating a completed Hardcopy data collection form. In other instances, the Soundex data are not manually collected. Rather, these data are calculated at time of entry of the surname data.



The process continues with a blank Electronic form/data entry screen used to capture Soundex:



In instances in which the Soundex code is manually transcribed on the hardcopy form, the value from the hardcopy form is entered into the Electronic form/data entry screen and then the edits and validations are performed on Soundex. In other instances, the Soundex code may be automatically calculated, based on the surname data:



The completed Electronic form/data entry screen is redisplayed and Soundex is stored in the database:

The screenshot shows a window titled "Electronic Form" with a blue header bar. Inside the window, there are several text input fields arranged in two rows. The first row contains "Prefix:" with the value "Mr.", "Last:" with the value "ANDERSON", and "Soundex:" with the value "A536". The second row contains "First:" with the value "JACKSON", "Middle:" with the value "ROBERT", and "Preferred:" with the value "JACK". Below the "Preferred:" field is a "Maiden:" field which is empty. To the right of the form fields, there is a large white arrow pointing from the "Soundex:" field towards a database storage area. The database storage area is represented by a cylinder icon and a box labeled "Database Storage". Inside the "Database Storage" box, the following information is displayed: "Variable: SOUNDEX", "Type: character", "Length: 4", and "Stored Value: A536".

Database Storage	
Variable:	SOUNDEX
Type:	character
Length:	4
Stored Value:	A536

Data Transmission: Electronic Data Interchange

Note: EDI sections are under construction.